NET+ Globus Service Evaluation

Product Description

Globus provides data sharing, transfer, compute, workflow, and security capabilities. Globus is Software-as-a-Service and Platform-as-a-Service and enables data movement from distributed storage endpoints. The service runs on AWS and presents RESTful APIs that are accessible via a web application, a command line interface, and via any application that can communicate over HTTPS. Data is never stored in Globus and does not flow through Globus; data always remains on endpoints (storage systems) managed by the institution.

Evaluation Group

The Functional Evaluation team was initially formed and charged with evaluating and testing Globus within the context of an Internet2 Request For Information (RFI) with the goal to find products that may be used to migrate data from Google Drive to other storage destinations. Globus is an important tool for managing data, and the evaluation speaks to both the Google Drive use-case and the tool's broad enablement of data mobility and sharing across storage systems (cloud and on-prem). The team consisted of 11 members across seven institutions, Internet2, and Globus.

Evaluation Team Members

Name	**Institution**
Dan Haven	University of Pittsburgh
Jim Leous	Penn State
John Entrup	Cornell University
Luis O. Hernández Muñiz	University of California, Berkeley
Ian Crew	University of California, Berkeley
Sam Porter	University of Maryland
Kevin Hildebrand	University of Maryland
John Regan	University of Maryland
Jon Mitchell	University of Washington
Jason Daughtrey	Indiana University
Charlie McClary	Indiana University
Vas Vasiliadis	Globus
Rachana Ananthakrishnan	Globus

Globus Concepts and Features

Endpoint

In v5, a single endpoint may include multiple Data Transfer Nodes or servers (each referred to as a Node); the endpoint provides the interface for server management and configuration.

Collection

Collections provide the interfaces required to access data. A collection consists of metadata about the collection, a DNS domain for accessing data on the collection, and configuration on Data Transfer Nodes that allows access to the data. There are two types of collections: Mapped (for users with local accounts on the storage system) and Guest (for users without local accounts, and often from external institutions).

Globus Connect Personal

Globus Connect is the software used to turn a system into an endpoint. Globus Connect Personal is used to create an endpoint on a single-user system such as a laptop or PC, allowing transfers to and from a personal machine.

Globus Connect Server

Globus Connect Server is used to create an endpoint on a multi-user system like a cluster or lab server.

Storage Connectors

Globus Connectors enable a uniform interface for accessing, moving, and sharing across storage systems. Globus Connect Server supports POSIX-compliance file systems, while Connectors enable other types of storage systems, for example Google Drive, Microsoft OneDrive, AWS S3, Box, CEPH, etc.

Globus Flows

Flows are a series of tasks that are executed in a specified order, and the Globus Flows service provides managed and reliable execution of flows. Each step in the flow is associated with an action provider that can act on a Globus resource or on externally managed services. Globus provides a number of action providers, and end users can create custom action providers.

Globus Compute

Globus Compute is a distributed Function as a Service Platform that allows remote function execution. Users launch persistent Globus Compute Endpoints on a system such as a laptop, campus cluster login node, cloud instance, or Kubernetes cluster, and use the Globus Compute Client for registration, execution, and management of functions across endpoints.

High Assurance

High Assurance is a set of features enabled on endpoints via a High Assurance or HIPAA BAA subscription, and is the product aimed at sensitive data with high security requirements. These features include higher authentication assurance for data access, isolation of applications and devices, forced encryption during transit, and more fine-grained audit logging.

Globus Management Console

The management console provides an interface to monitor and manage activity on managed endpoints and managed collections.

Evaluation Methodology

The team evaluated Globus responses to 50 use cases encompassing features across functionality, security, identity, and administration by drawing upon historical experiences within the group, discussion with Globus team members, and hands-on testing with Globus subscriptions. The use cases were developed as part of the Internet2 RFI, and the team added additional use cases that were not in the RFI. The team also dedicated a portion of the evaluation to review experiences related to the use of Globus as a tool for migrating data from Google Drive to another storage destination.

Results of Testing and Use Case Review

Product Onboarding Experience

The Globus onboarding experience includes a link to a Welcome Kit that includes documentation and videos describing Globus setup and configuration. There is also a strong focus on how to match use cases and increase usage of Globus once it has been deployed at an institution. Globus also offers on-demand consultations for configuration and setup, and provides support after the product has been deployed.

Functional

Globus is used by thousands of institutions on a regular basis for file transfers and features a variety of functional capabilities.

Our testing generally indicated that Globus performed as described, and no Globus issues were encountered while installing and configuring Globus Connect Server and Connectors, including Standard and High Assurance editions.

For this evaluation, we reviewed 28 use cases related to Globus functionality within the context of data migrations across different storage systems.

- Support for Administrator migration - Globus is not a migration tool, but data transfers on behalf of end-users could be enabled with applications using Globus APIs
- Self-service tools for data migration - Globus is an end-user focused data transfer tool; self-service is possible.
- Manage migration through command line - All data transfer actions are possible through CLI
- Manage migration through web interface - Data transfer may be initiated via web application

- Data storage location during migration - Data is not stored or cached and glows over a data channel between the source and destination storage services
- Move data larger than 25TB - There is no limit to the amount of data that can be transferred
- Move files greater than 250GB - There is no limit to the size of files that can be moved
- Capable of moving thousands of accounts / TB per day - Globus is not a migration tool. Instead, it enables data transfers for end-users and is primarily limited by source and destination storage systems.
- Average data rates and limiting factors - Globus provided these observed data rates:
    - Google Drive to Google Cloud Storage: 800GB dataset comprising 400 x 2GB files; effective speed averaged ~3.6Gbps.
    - Box to Google Cloud Storage: 1TB and 3TB datasets comprising a mix of 100MB, 1GB files and 2GB files; effective speed averaged ~3.2Gbps.
    - On-premises HPC storage to Google Cloud Storage: 1TB dataset comprising 100 x 10GB files; effective speed averaged ~6Gbps.
    - The primary constraint in the cloud-to-cloud cases are the API rate limits (set on individual accounts by the provider) which downgrade overall throughput.
- Google native file handling - Native files are not migrated and files are moved in native formats
- Move data and delete at source capability - Possible with the use of a Globus Data Flow
- Ability to choose file types to move - Possible using a list of source-destination pairs of file names as a batch to be transferred via the Globus command line interface.
- Ability to ignore certain file types - Possible to exclude files and directories using pattern matching
- Ability to move data in Shared Drives - Data in Shared Drives can be transferred
- Data transfer interruptions - Robust monitoring and control mechanism in place to pause and resume transfers, including backoffs related to API limits in source/destination storage services
- Target destination storage capacity - Not a feature
- Data transfer interruptions requiring end-user intervention - Transfer resumes when rectified
- Transfer making no progress - End user may specify a deadline for action if no progress is made
- Automatic issue remediation capabilities - A variety of errors are automatically addressed and certain types of errors may be set to be ignored.
- Data integrity verification - Checksums are used to verify integrity; re-queued if failure.
- Failed transfer handling - User is notified about failures via email
- Pre-migrate data - Globus is not a migration tool but multi-step data transfers are supported with Globus Flows
- Preservation of source sharing and permissions - Globus does not automatically transfer sharing settings
- Re-share migrated data with specific people internal / external to Globus - permissions are not inherited; ACLs would need to be duplicated manually or via script/program; Globus guest collections may also be used for sharing
- Notification type and frequency for internal users when re-sharing data - Notifications are sent when data is initially shared. Frequency is not customizable.
- Notification type and frequency for external users when re-sharing data - Notifications are sent when data is initially shared. Frequency is not customizable.
- Notifications for migration start - Not by default
- Notifications for migration end - Sent to end-user
- Migration report - Summary and detailed information is provided to the end-user, and may be accessed manually.
- Set Google Quotas after migration - Globus does not set quotas on storage systems.
- Set Google Quotas on Shared Drives after migration - Globus does not set quotas on storage systems

- End-user pay options for migration - Globus is only available at the institution level
- Log Availability for specific user transfers - Real time transfer details are available to administrators via the management console. Transfer details are available to the user on the Activity page. Summary level information for all transfers on managed endpoints are available to authorized users in the usage reporting collection.

Identity

Globus allows user authentication through many supported identity providers (ORCID, Google, institutional identities). Globus acts as an identity broker and uses federated login; institutional credentials are never seen by Globus. A Globus Account is the set of linked identities used to log in to Globus. Globus identities are used to authorize access to data, and several mechanisms are available to control and limit how access can be granted. Because Globus links identities to a single account, it is possible to share data easily with collaborators across multiple organizations. In addition, Globus Groups may be used to share files and folders with multiple collaborators at once, and/or to manage permissions in one location for multiple files and folders.

We evaluated specific two use cases for identity.

- Limit who can perform self-migrations to a specific group or OU - May be accomplished by using the User Policy (--user-allow and --user-deny) to limit access to the storage gateway for Google Drive, or by using Globus Groups and managing Group membership.
- Preventing ‘linked identity’ access to a collection / endpoint - This requires High Assurance and admin configuration that restricts access to specific domains (Authentication Policy --domain).

Security and Compliance

Globus has completed the Higher Education Community Vendor Assessment Toolkit (HECVAT) and maintains a Business Associates Agreement (BAA) with AWS. If a subscriber institution intends to use the High Assurance subscription to move data subject to HIPAA, Globus provides a BAA as an add-on to your subscription. Globus inherits physical controls, and many other controls, from AWS where the service is hosted. In addition, Globus has many practices and procedures in place to secure the Globus infrastructure. Globus indicated that Per the HECVAT, Globus maintains internally documented Change Control Procedures, Cryptography Policies and Procedures, Contingency Plans, Hardware Usage, Incident Response Plans, Technical Access Controls, and Server and Cloud Monitoring Plans.

Indiana University completed a review of Globus security practices and determined that it presents low risk to the institution. In particular, Globus does no intermediate data caching and Globus infrastructure is able to inherit security controls instituted by AWS. **Primary risks result from file sharing practices by end-users sharing files with the wrong collaborators, giving access to the wrong folders, giving overly permissive permissions, and file deletion activities resulting from read-write permissions.**

The High Assurance (HA) subscription introduces security controls that are critical for protecting sensitive data including re-authentication requirements, default encrypted data transfers, and audit logging. In particular, HA allows admins to set policy that requires end users to authenticate with specific identities (rather than a linked identity), require authentication for new sessions and devices, and timeout authentications after a specific period of time.

We evaluated nine specific use-cases related to Security, using our collective experience and testing to verify if the use cases are met by Globus.

- Encryption - Globus is capable of encrypting data transfers
    - Globus does not store data; Data is moved between endpoints using a “data channel” that is only accessible to the GridFTP servers running on the endpoints. The channel is authenticated by default, but unencrypted. Encryption is an option available to the end-users through the File Manager of CLI, and by default the OpenSSL cipher is AES256-SHA.
- Data Retention - Globus does not store or retain data; but does store/retain some logs.
    - The Globus service only stores control data to manage file transfers and ensure their successful completion while maintaining data integrity. Such control data (e.g., file/object names) is stored temporarily by Globus on AWS and is either destroyed after a specified period or archived for audit purposes, depending on the storage system(s) involved in the transfer.
- Globus access token revocation
    - Globus has no access to identity provider credentials; instead Globus access tokens can be explicitly revoked (e.g., to remove access to a mapped collection). This is typically not required for many scenarios because the token is evaluated at time of use, and if an end user no longer has access to the underlying storage system (e.g., because they were terminated) then access will fail, even if the token has not yet expired.
    - In addition, admins have controls such as pausing in-progress transfers on an endpoint.
- Source endpoint file ownership and permissions handling - Globus will not copy ownership/sharing permissions from source destinations.
    - Ownership and permissions are not automatically maintained. Ownership of the destination files is set to the user account that requested the transfer. Permissions are determined entirely by the destination Globus endpoint’s configuration. You can control the permissions of the files created by Globus, on a destination Globus endpoint, but it is not based on information about the original file permissions. Permissions may be updated by a script that initiates the transfer via the Globus command line interface and then sets the appropriate permissions on the destination files.
- Source endpoint security controls handling - Globus performs actions as the user and will be limited to permissions provided by the source endpoint.
    - All operations are performed as the user (mapped user or guest owner), so after GCS login and path restrictions/ACL checks, it is entirely up to the underlying storage system (Google Drive, Box, etc.) whether that user's credential is allowed to do something.
- Admin/end-user ability to limit access to content - Access may only be limited based on file name patterns; end-users need only read access to copy files using Globus.
    - The administrator/end user may transfer any files to which they have (at least) read access. Exclusion rules may be specified based on file name patterns, not on permissions.
- Flexible logging - log availability depends on the action and subscription being used.
    - Multiple logging levels are available and logs are written at varying levels of granularity, depending on the type of subscription that institution has.
- Splunk Integration
    - Logs are written by Globus Connect directly to the subscriber’s storage system. Management of the logs, such as policies and procedures for access, encryption, and retention, are the responsibility of the subscriber
- Logs viewable by connector - A subscriber would need to aggregate usage of corresponding collections in order to view usage by connector type

    - GridFTP has a logging option to send usage stats to a receiver. These usage stats can be configured to include the connector type used for the transfer. The subscriber will need to configure a receiver and roll the logs up into aggregate usage information.

Accessibility

Globus has provided the Voluntary Product Accessibility Template (VPAT) and Luis O Hernández Muñiz has performed the accessibility review. Overall, Globus is in great shape and there's only one item to address. The Globus team self-reported that there are some bugs in the keyboard navigation of the Globus Connect Personal preferences pane. Globus plans to comprehensively test the keyboard navigation of the preferences pane, define the scope of work, and provide a remediation timeline. That should be included in the agreement moving forward unless they have already addressed it.

Network

Only one use-case was identified related to the network to describe the network path used between Google Drive and destination endpoints. Transfers will use any available route between source and destination systems. A Globus endpoint may be configured to use a specific network interface that will route data traffic over a specific path.

Integration

Three use cases related to Integration.

- List Supported storage systems - see below
    - Google Drive (including Shared Drives)
    - Google Cloud Storage
    - Microsoft OneDrive (including SharePoint sites)
    - Microsoft Azure Blob Storage
    - Box
    - Amazon Web Services S3
    - Any fully S3-compatible storage (both public-cloud and on-premises)
    - OpenStack Ceph RADOS Gateway
    - Spectra Logic BlackPearl
    - Quantum ActiveScale
    - iRODS
    - HPSS
    - HDFS
    - (Support for Dropbox is under development)
- Migrations from other platforms - Globus is not a ‘migration’ tool but may be used to transfer data between any of the supported storage systems.
- API availability for integrations - Globus services provide a REST API using security based on the OAUTH framework.

Administration

In the Globus Platform, system administrators primarily create, deploy and manage endpoints with Globus Connect Server. Administrators are also able to view various logs (standard GridFTP logs for base subscription and GridFTP audit logs with High Assurance Endpoint Collections.

Typical administrator responsibilities include configuring Storage Gateway Connectors, Authentication Requirements and Policies, Identity Mapping policies, Data Access Policies, Use Access Restrictions, creating Collections, and setting metadata.

The Globus Connect Server CLI and API support role based authorization so that administrators can delegate ability to perform administration tasks on an endpoint or a collection to others. These roles may be associated with either a Globus Auth user identity or with a globus group, which grants that role to all members of that group. These roles are Owner, Administrator, activity_manager and activity_monitor.

Globus for Google Drive Migrations

During our second functional review meeting, a participating institution shared their experience using Globus as a tool for end-user data migrations from Google Drive to other storage services. Globus is not a data migration tool intended for large-scale, multi-account data transfers; it is intended to be used by end-users to move their own data between a variety of endpoints. **Globus was not the only migration tool used by the institution, but was a key part of the toolbox available to end-users and admins during the migration.** Without a tool like Globus to enable extremely large data transfers across numerous storage services with a high level of reliability for success and reporting, the institution would likely have lost TBs of data that had no other reasonable path for migration. Institutions moving all of their Google Drive accounts and data may find that other tools are necessary for a successful migration.

Globus was most often used when the storage destination would not be Microsoft OneDrive, and was used by end users to move data from My Drive and Shared Drive locations. Because of the way Globus manages permissions to endpoints, there is not an Administrative migration functionality; Globus is an end-user focused tool. However, admins are able to access reports and logs for end-user Globus activity in order to assist with troubleshooting failed transfers. Data Transfers also result in summary reports emailed to end-users with information about file transfer success and failures. Other notable considerations for using Globus for data transfer out of Google Drive:

- Google Format files (docs, sheets, slides) do not transfer[1]
- Large numbers of files will likely be throttled by Google API limits; Globus will automatically adjust to ensure successful transfer but duration of the move is greatly increased. There may also be API limits on the storage destination endpoint that impact duration. [2]
- Data verification issues were experienced with certain Microsoft files (pptx) as a result of random changes to checksum values; the issue was determined to be on the Microsoft side and a resolution is to ignore the verification failure and transfer anyway, or to manually move the failed files.
- Globus cannot process multiple objects with the same name in the same directory; the files will need to be deleted or renamed. [3]
- Globus Connectors are required for endpoints like Google Drive, Microsoft OneDrive, etc.
- Globus support was contacted a few times during the migration. Responses were timely and helpful. The issues turned out not to be related to Globus.

[1] This limitation was discussed at length as part of the evaluation and will be explored as a future capability by Globus.
[2] It was noted during the evaluation that additional support information would be helpful; Globus has added this need to their documentation backlog
[3] This limitation was discussed at length as part of the evaluation and will be explored as a future capability by Globus.